DATAWAREHOUSING AND MINING

UNIT-1

1.1 INTRODUCTION TO DATA MINING

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. With the widespread use of databases and the explosive growth in their size organizations are faced with the problem of information overload. The problem of effectively utilizing these massive volumes of data is becoming a major problem for all enterprises. Data mining techniques support automatic exploration of data and attempts to source out patterns and trends in the data and also infer rules from these patterns which will help the user to support review and examine decisions in some related business or scientific area.

Need for data mining?

The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format .This accumulation of data has taken place as at an explosive rate.

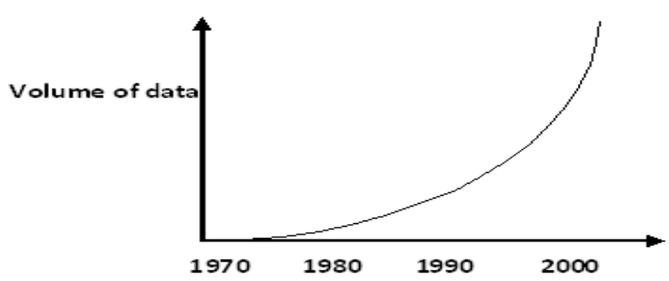


Fig: 1.1 the growing base of data

Data storage became easier as the availability of large amount of computing power at low cost i.e. the cost of processing power and storage is falling, made data cheap. There was also the introduction of new machine learning methods for knowledge representation based on logic programming etc, in addition to traditional statistical analysis of data. The new methods tend to be computationally intensive hence a demand for more processing power.

Having concentrated so much attention on the accumulation of data the problem was what to do with valuable resource? It was recognized that information is at heart of business operation and that decision-makers could make use of the data stored to gain valuable insight into the business. Database management systems gave access to the data stored but this was only a small part of what could be gained for the data stored to gain valuable insight to the business.

Database management system gave access to the data stored but this was only a small part of what could be gained from the data. Traditional on-line transaction procession systems, OLTPs, are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return.

Analyzing data can provide further knowledge about the business by going beyond the data explicitly stored to drive knowledge about the business. This is where data mining or knowledge discovery in database (KDD) has obvious benefit for any enterprise.

WHAT IS DATA MINING?

Data mining refers to extraction or mining knowledge from large data bases. Data mining and knowledge discovery in the database is a new interdisciplinary field, merging ideas from

statics, machine learning, databases and parallel computing.

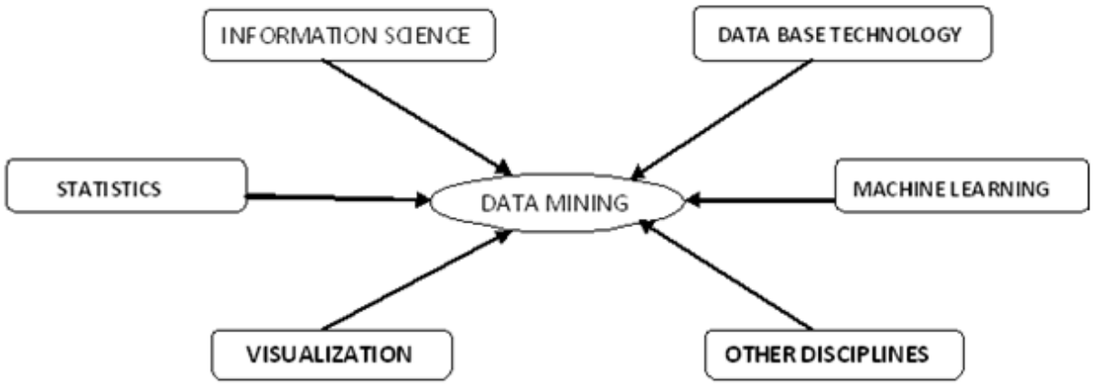


Figure 1.2 *Data mining as a confluence of multiple Disciplines*

Some of the definitions of data mining are:

1. Data Mining is the non –trival extraction of implicit, previously unknown and potentially useful information from the data.
2. Data Mining is the search for the relationship and global patterns that exist in large databases but are hidden among vast amounts of data.
3. Data Mining refers to using variety of techniques to identify nuggets of information or decision-making knowledge in the database and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting and estimation .
4. Data mining system self learn from the previous history of investigated system, formulating and testing hypothesis about rules which system obey.

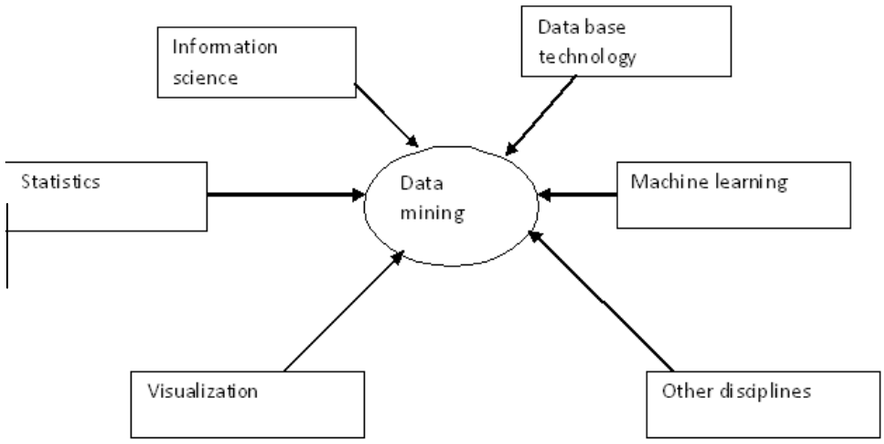
5. Data mining is the processer of discovering meaningful new correlation pattern and trends by shifting through large amount of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques.

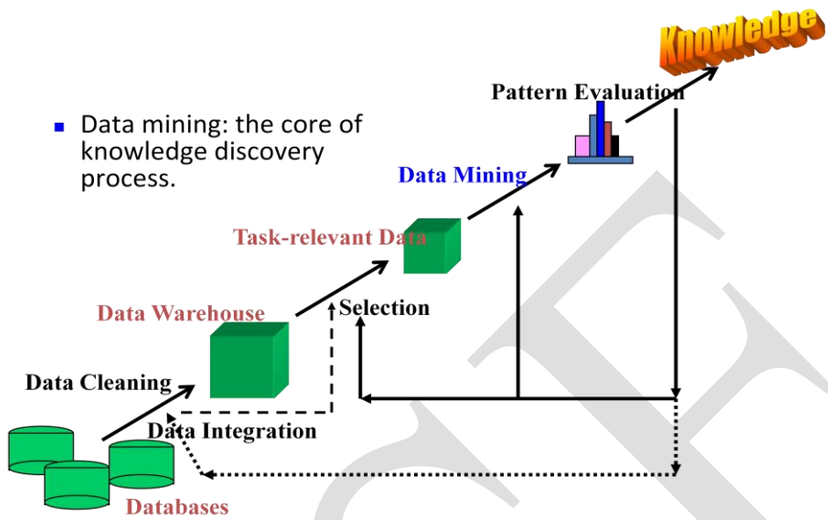
KDD vs. Data mining.

Knowledge Discovery database (KDD) was formalized 1989, with reference to general concept of being broad and high level in the pursuit of seeking knowledge from data. Data Mining is only one of the many steps involved in knowledge discovery in data bases .The KDD process tends to be highly iterative and interactive. Data mining analysis tends to work up from the data and the best techniques are developed with an orientation towards large volumes of data, making use of as much data as possible to arrive at reliable conclusions and decisions. Fayyad et.al distinguishes between KDD and data mining by giving the following definitions:

Knowledge Discovery in Database (KDD) is the process of identifying a valid, potential useful and ultimately understandable structure in data. This process involves selecting or sampling data from a data warehouse, cleaning or preprocessing it, transforming or reducing it applying a data mining component to produce a structure, and then evaluating the derive structure

Data mining is a step in the KDD process concerned with the algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency limitations. patterns or structures are enumerated from the data under acceptable computational efficiency limitations.





- Data mining: the core of knowledge discovery process.

Data Cleaning: It is the process of removing noise and inconsistent data.
Data Integrating: It is process of combining data from multiple sources.
Data Selection: It is the process of retrieving relevant data from database.
Data Transformation: in this process, data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
Data Mining: it is an essential process where intelligent methods are applied in order to extract data patterns.
Pattern Evaluation: The patterns obtained in the data mining stage are converted into Knowledge based on some interestingness measures.
Knowledge Presentation: Visualization and Knowledge representation techniques are used to present the mined knowledge to the user.

1.2. AREHITECTURE OF DATA MINING SYSTEM

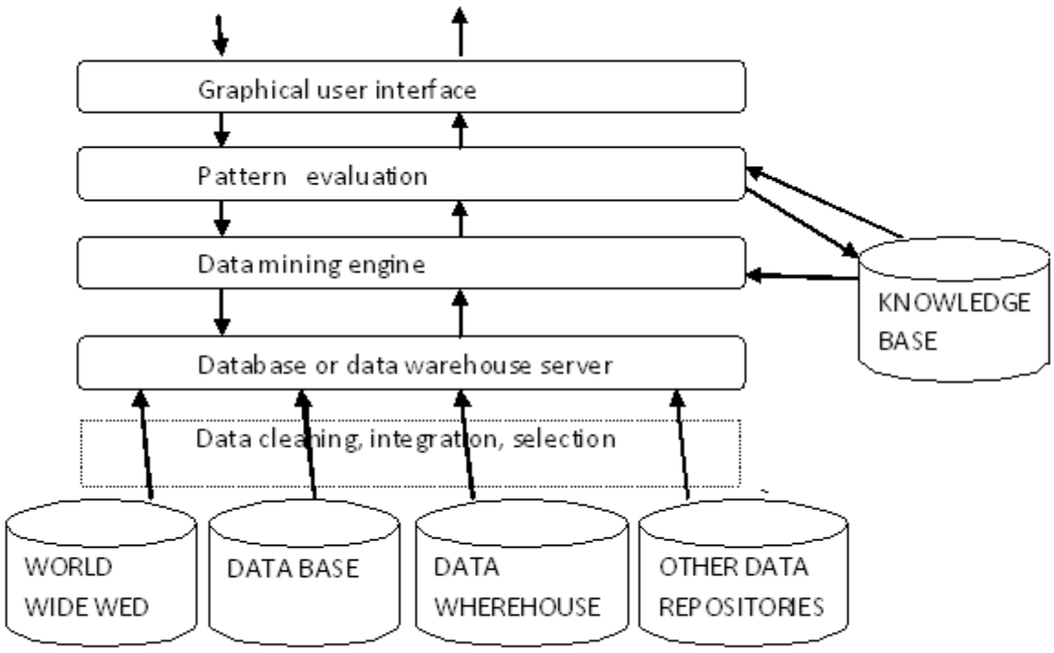


FIGURE 1.4 Architecture of a typical data mining system

Data mining is the process of discovering interesting Knowledge from large amounts of data stored either in databases, data warehouse or other information repositories based on this view, the architecture if a typical data mining system may have the following major components.

*Database, Data Warehouse or Other Information Repository: This is a single or a collection of multiple databases, data warehouses, Flat files, spreadsheets or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data

*Database or data warehouse server: The database or data warehouse server fetches the relevant data, based on the user's data mining request.

*knowledge Base: This is the domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attribute values into different levels of abstraction knowledge such as user beliefs, thresholds and metadata can be used to access a pattern's interestingness.

*Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for task such as characterization, association, classification, cluster analysis, evolution and outlier analysis.

*Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards increasing patterns. It may use interestingness thresholds to filter out discovered patterns. Alternately, the pattern evaluation module may also be integrated with mining module.

*Graphical user interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a task or data mining query for performing exploratory data mining based on intermediate data mining results. This module also allows the user to browse database and data ware house schemes or data structures, evaluate mined patterns and visualize the pattern in different forms such as maps, charts etc.

1.3 Data mining- on what kind of data?

Data mining can be applied to any kind of information repositories such as relational databases, data warehouse, transactional data bases, advanced systems, flat files and the World Wide Web. Advanced databases system include object- oriented and object – relation databases, and specific application – oriented data bases such as spatial data base, time series data bases text databases and multimedia databases.

1.3.1 Relational databases

A relational database is a collection of tables. Each table consists of a set of attributes (columns or field) and a set of topples (records or rows). Each topple is identified by a unique key and is described by a set of attribute values. Entity relationship (ER) data model is often constructed for relational databases. Relational data is accessed by database queries written in a relational query language.

Employee

Emp-id	Name	Dept	Salary
E25	Ram	Account	10,000

1.3.2 Data Warehouse

A data Warehouse is a repository of information collected from multiple sources, stored under a unified scheme residing on a single site.A data warehouse is modeled by a multidimensional database structure, where each dimension to an attribute or a set of attributes in the schema.

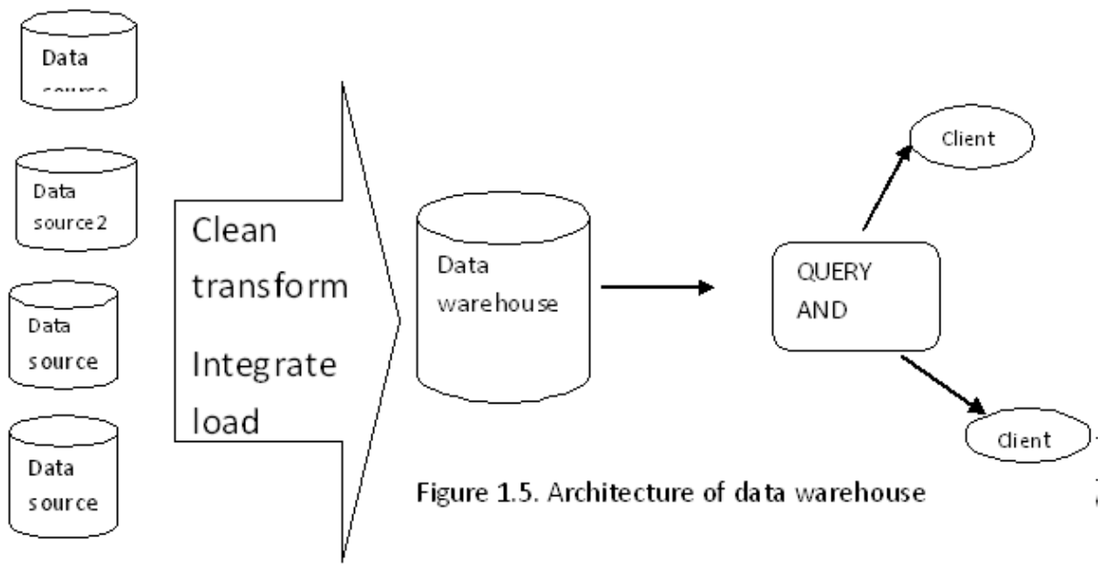


Figure 1.5. Architecture of data warehouse

Data ware house is modeled by data cubes. Each dimension is an attribute and each cell represents the aggregate measure. A data ware house collects information about subjects that span an entire organization whereas data mart focuses on selected subjects. The multidimensional data views marks (OLAP) online analytical processing easier.

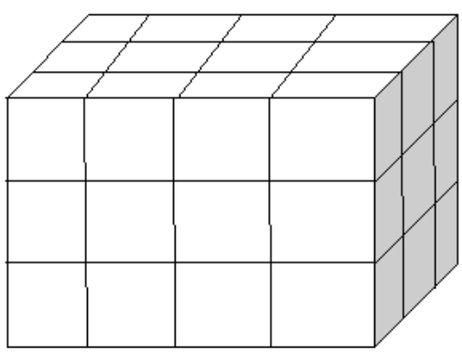


Figure 1.6 Data Cube

1.3.3 Transactional databases

A transactional database consists of a file where each record represents a transaction. A transaction includes transaction identity number, list of items, date of transactions etc.

Sales

Trans_ ID	List of items	Trans_ data
T100	I1,I5,I6	15-06-2007

1.3.4 Advanced databases

Object oriented databases

Object oriented databases are based on object – oriented programming concept. Each entity is considered as an object which encapsulates data and code into a single unit. Objects are grouped into a class.

Object – Relational data bases

Object relational databases are constructed based on an object relational data modal which extends the basic relational data modal by handling complex data types, class hierarchies and object inheritance.

Spatial Databases

A spatial database stores a large amount of space – related data, such as maps, preprocessed remote sensing or medical imaging data and VLSI chip layout data. Spatial data may be represented in raster format, consisting of n-dimensional bit maps or pixel maps.

Temporal databases and time – series databases

Temporal databases and time series databases both store time relation data. A temporal database usually stores relational data that include time – related attributes. A time series database consists of sequence of values or events changing with time.

Text databases and multimedia databases

Text databases contains word descriptions for objects such as long sentences or paragraphs, warning messages, summary reports etc.Text database consists of large collection of documents from various sources. Data stored in most text databases are semi structured data. A multimedia database stores and manages a large collection of multimedia objects such as audio data, image, video, sequence and hypertext data.

Heterogeneous Databases and legacy databases

A legacy database is a group of heterogeneous databases that combines different kinds of data systems.

The World Wide Web

The World Wide Web is a popular and interactive medium to disseminate information today. The web is huge, diverse and dynamic and thus raisers the scalability, multimedia data and temporal issues respectively.

1.4. DATA MINING FUNCTIONALITIES

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks are classified into two categories descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

1.4.1 Concept / Class Description

Descriptions of a individual classes or a concepts in summarized, concise and precise terms called class/ concept descriptions. These descriptions can be divided via

1. Data Characterization
2. Data Discrimination
3. Both data characterization and discrimination.

Data characterization:

- It is a summarization of the general characteristics of a target class of data.
- The data corresponding to the user specified class are collected by a database query.

Several methods like OLAP roll up operation and attribute-oriented technique are used for effective data summarization and characterization.

The output of data characterization can be presented in various forms like

- Pie charts
- Bar charts
- Curves
- Multimedimensional cubes
- Multimedimensional tables etc.

The resulting descriptions can be presented as generalized relations or in rule forms called characteristic rules.

Data Discriminations:

Comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

The output of data discrimination can be presented in the same manner as data characterization. Discrimination descriptions expressed in rule form are referred to as discriminate rules

1.4.2 ASSOCIATION ANALYSIS

Association analysis is the discovery of association rules showing attribute value conditions that occur frequently together the given set of data.

An association rule is an expression of the form X ⇨ Y, where X and Y are the sets of items. This rule implies that the transaction of the data base which contains X tends to contain Y this rule should satisfy two interesting measure namely support and confined.

Given a database, the goal is to discover all the rules that have the support and confined greater then or equal to the minimum support and confidence, respectively.

Support means how often X and Y occur together as a percentage of the total transaction.

Confidence measures how much a particular item is dependent on another

Association rules that contain a single predicate are referred to a single dimensional association rule.

Ex: Buys (X,"computer") => Buys (X,"software")

Association between more than one attribute or predicate is referred as multidimensional association rule.

Ex: age (X, "20……..28") ^income (X,"25k…….40k") => buys (X,"computer")

1.4.3 Classification and prediction

Classification involves finding rules that partition the data disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classification analyzes the training data set and constructs a modal based on the class label and aims to assign a class label to the future unlabeled records. There are several classifications models, they are decision tree, neural networks, Genetic algorithms, mathematical formulae like liner / geometric discriminates etc...

Prediction is used to predict missing or unavailable data values rather than class labels. Prediction refers to both data value prediction and class label prediction. The prediction of continuous values can be modeled by statically techniques of regression.

1.4.4 *Cluster Analysis*

Clustering is a method of grouping data into different groups, so that in each group share similar trends and patterns. The objectives of clustering are:

)-(To uncover natural groupings.
)-(To initiate hypothesis about the data
)-(To find consistent and valid organization of data.

Clustering analyzes data objects without considering known class label. Clusters can be grouped based on the principles of maximizing intra- class similarly and minimizing inter class similarity.

1.4.5 *Outlier Analysis*

Data objects which differ significantly from the reaming data objects are referred to as outliers. Normally outliers do not comply with the general behavior or model of the data. Hence, most of the data mining methods discard outliers as noise or exceptions. Outliers mining are used for identifying exceptions or rare events which can often lead to the discovery of interesting and unexpected knowledge in areas such as credit card fraud detection, cellular phone cloning fraud and detection of suspicious activities. Some of the techniques for detecting outliers are statically test, distance measures and deviation based method.

1.4.6 Evolution Analysis

Data Evolution analysis describes and model regularities (or) trends of objects whose behavior changes over time. Normally, evolution analysis is used to predict the future trends by effective decision making process.

1.5 CLASSIFICATION OF DATA MINING SYSTEM

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including data base systems, statistics, machine learning, and visualization and information source. Data mining system can be classified according to the kinds of databases mined, types, techniques utilized, functionalities applied etc.

- Classification according to the kinds of databases mined

Database systems themselves can be classified according to different criteria such as data models, or the of data or application involved.

(i) Classification according to data models

)-(Relational
)-(Transactional
)-(Object oriented
)-(Object relational
)-(Data warehouse mining system

(ii) Classification according to special types of data handled

)-(Spatial data mining
)-(time series data mining

)-(Text or multimedia data mining
)-(WWW mining system

- Classification according to the kinds of knowledge mined

(i) Based on data mining functionalists

)-(Characterization, dissemination
)-(Association, classification
)-(Evolution analysis.

(ii) Based on granute dfrity *or levels of abstraction.*

)-(Generalized knowledge
)-(Primitive- level knowledge.
)-(Knowledge at multiple levels.

(iii) Based on regularities and irregularities

)-(Commonly occurring patters
)-(Expectation or outliers

- Classification according to the kinds of techniques utilized

(i) According to the degree of interaction

)-(Autonomous data mining system
)-(Query Driven systems
)-(Interaction exploring systems

(ii) Based on methods of data analysis

)-(Data base oriented
)-(Data warehouse oriented
)-(Machine learning
)-(statistics
)-(visualization
)-(pattern reorganization
)-(Natural network

- Classification according to the applications, adapted such as finance, telecommunications, DNA, stock markets, banking, retail. E- Mail and so on.

1.6 DATA MINING TECHNIQUES

The most commonly used techniques in data mining are:

Neural Networks:

Neural Networks have the remarkable ability to derive meaning from complicated or imprecise data can be used to extract pattern and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neutral network can be thought of as an "expert "in the category of information it has been given to analyze. These experts can be used to provide projections given new situation s of interest and answer "what if" questions.

Neutral networks use a set of processing elements a (or nodes) analogous to neurons in the brain. These processing elements are interconnected in a net work that can be identify patterns in data once it is exposed to data, i.e., the network learns from experience just as people

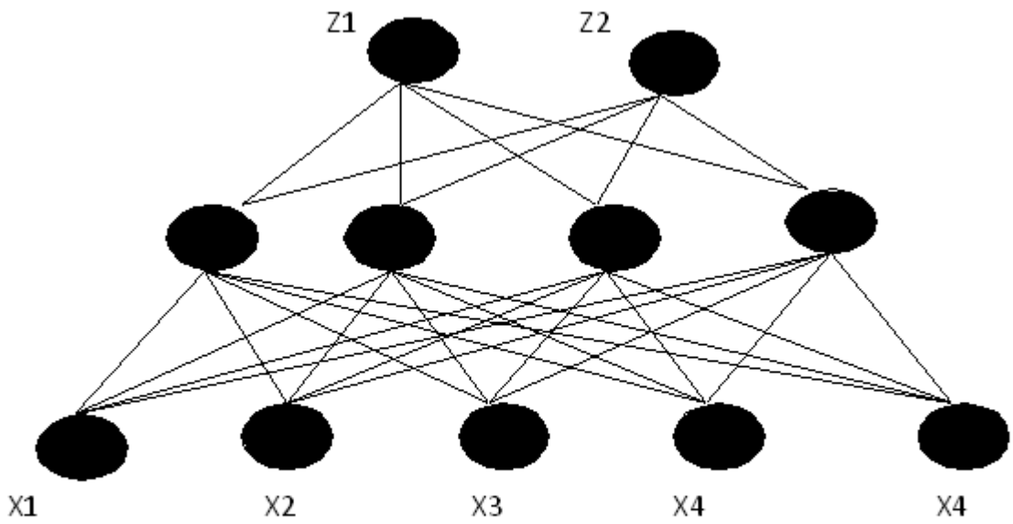


Figure 1.7 structure of a neural network

do.

The bottom layer represents the input layer, in this case with 5 input label X1 through X5 .in the middle is something called the hidden layer, with a variable number of nodes .The out put layer in this case has two nodes, Z1 and Z2 representing out put values we are trying to determine from the inputs. Each node in the hidden layer I s fully connected to the inputs which means that what is learned in a hidden node is based on all the inputs that taken together...

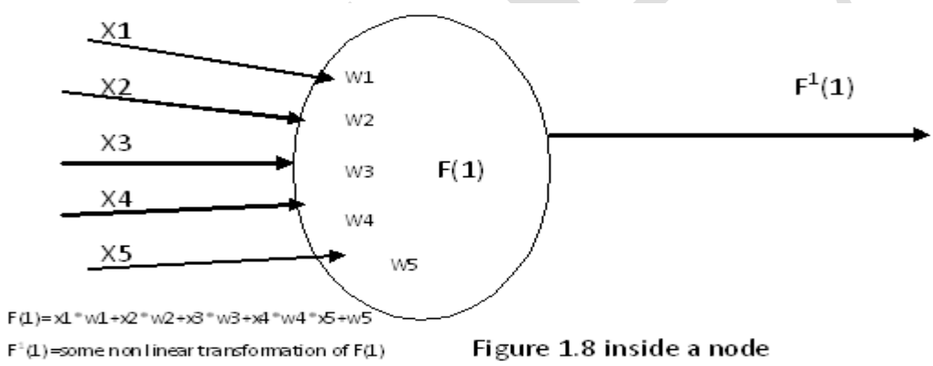


Figure 1.8 inside a node

- *Decision trees*

  Decision trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes are labeled with attribute names, the edge are labeled with possible values for this attributes and leaves labeled with different classes. Tree shaped structure s represents sets of decisions. These decisions generate rules for the classification of a dataset. Decision trees produce rules that are mutually exclusive and collectively exhaustive with respect to the training databases .Specification decision tree method includes classification and regression trees (CART) and chi square automatic interaction detection (CHAID). The following is an example of object that describes the weather at a given time .The objects contain information on the outlook, humidity etc. Some objects are positive examples denoted by P and other are negative i.e. N.

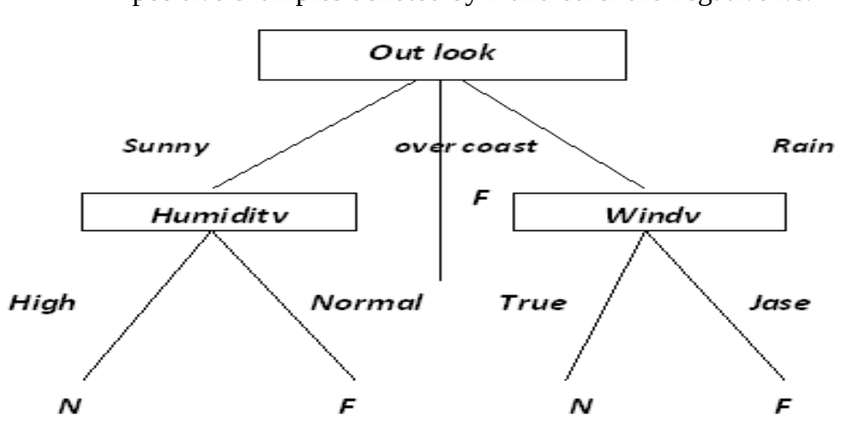
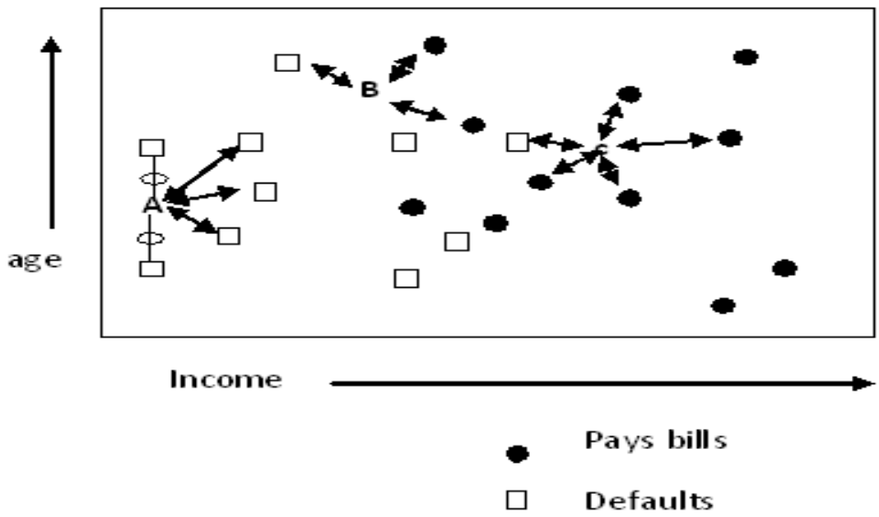


Fig:1.9 Decision tree structure

Nearest neighbor Method: A technique that classifies each record in a data set based on a combination of the classes of the K record(S) most similar to it in a historical dataset (whereK^3 1) is cacibed in terms of measurements or by relationship with other objects. Clustering is sometimes used to mean segmentation. Clustering and segmentation basically partition the database so that each partition or group is similar according to some criteria or metric. Many data mining applications make use of clustering to similarity for example to segment a client/customer base. Some of the clustering algorithms are DBSCAN, CHAMELEON and k-



medics.

Rule induction

Rule induction is the process of extracting useful if –then rules from data based on statistical. rule induction on a data base can be a massive undertaking in which all possible pattern are systematically pulled out of the data and then accuracy and significance calculated, telling users how strong the pattern is and how likely it is to occur again

Genetic Algorithms

Genetic algorithms refer to the algorithm that dictates how populations of organisms should formed, evaluated and modified. Genetic algorithm is a optimization techniques that use processes such as genetic combination, mutation, and natural selection in a has a variety of forms, but in general their application is made on top of an existing data mining techniques such as neural net works or decision tree

Data visualization

Data visualization makes it possible for the analyst to gain a deeper, more intuitive understanding of the data and as such can work well along side data mining. Data mining allows the analyst to focus on certain patterns and trends and explore in-depth using visualization. On its own data visualization can be overwhelmed by the volume of data in a data base but in conjunction with data mining can help with exploration.

1.7 Major issues in data mining

Major issues in data mining are mining methodology, user interaction, performance, and diverse data types. These issues are introduced below:

Mining methodology and user interaction issues:

- *Mining different kinds of knowledge in databases:*

Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks which may use same database in different ways and require the development of numerous data mining techniques.

- *Interactive mining of knowledge at multiple levels of abstraction:*

  Interactive mining allows users to focus the search for patterns, providing and refini8ng data mining requests based on returned results to view data and discovered patterns at multiple granularities and from different angles.
- *Incorporation of background knowledge:*

Background knowledge or domain knowledge guides the discovery process in concise terms at different levels of abstraction and also speed up a data mining process, or judge the interesting of discovered patterns.

- *Data mining query languages and ad hoc data mining :*

  High-level data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns.
- *Presentation and visualization of data mining results:*

  Using visual representations, or other expressive forms, the knowledge can be easily understood and directly usable by humans. This requires the system to adopt expressive knowledge representation techniques, such as tree, tables, rules, graphs, charts, crosstabs, matrices, or curves.
- *Handling noisy or incomplete data:*

  Noise or incomplete data may be confuse the process, causing the knowledge model constructed to over fit the data which intern result the accuracy of the discovered patterns to be poor data clearing methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases.

- *Pattern evolution – the interestingness problem:*

  A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty. Several challenges 0remain regarding the development of techniques to asses the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measure or user specified constraints to guide the discovery process and reduce the screech space in another active area of research

  Performance issues:

  This include efficiency scalability, and parallelization of data mining algorithm
- Efficiency and scalability of data mining algorithms:

  in order to effectively extract for information from huge amount data in databases, data mining algorithms must be efficient and scalable as well as running time of a data mining algorithms must predictable and acceptable.
- Parallel, distributed, and incremental mining algorithms:

  The huge size of databases, wide distribution of data, high cost and computational complexity of data mining methods leads to the developments of parallel and distributed data mining algorithms moreover, the incremental data mining algorithms updates database without having to mine this entire data again "from scratch"

Issues relating to the diversity of database types

- Handling of relational and complex types of data:

It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining specific data mining systems should be constructed for mining specific kinds of data. Therefore, one may expect to have different data mining systems for different kinds of data.

- Mining information from heterogeneous databases and global information systems:

Data mining may help disclose high level data regularities in multiples heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases. Web mining, with uncovers interesting knowledge about web contents, web structures, web usage, and web dynamics, becomes a very challenging and fast evolving field in data mining.

Data mining systems relay on databases to supply raw data for input and this raises problems in that databases tend be dynamic, incomplete, noisy, and data large. Other problems arise as a result of the adequacy relevance of the information stored. The above issues are considered major requirements and challenges for the further evolution of data mining technology.